Viability Status of Oregon Salmon and Steelhead Populations in the Willamette and Lower Columbia Basins

Appendix A: Random Multinomial Finite Sampling Method

June 2007

Paul McElhany
NOAA-Northwest Fisheries Science Center

Prepared for
Oregon Department of Fish and Wildlife and
National Marine Fisheries Service

Note: This method is used to include uncertainty in our estimate of the population attribute weights for calculating overall population risk status and for adding error to age structure parameters in the viability curves.

In simulation modeling we often encounter parameters that are a vector of fractions partitioning some whole. For example, may have a parameter that describes the fraction of fish that spawn at a given age, with 20% at age 3, 70% at age 4 and 10% at age 5. The fractions need to sum to 100% and are therefore clearly not independent. Other examples are weights on some linear function where the weight totals must sum to one or the fraction of the habitat that is in specific categories. These fraction vectors are often estimated with uncertainty and for Monte Carlo simulations we need to randomly generate new vectors that sum to one and have a controlled distribution around the point estimates.

If we treat the fractions as probabilities, the vectors describe a multi-nominal distribution (e.g., the probability of age 3 is 0.2, the probability of age 4 is 0.7 and the probability of age 5 is 0.1). To obtain a random vector with the appropriate properties we apply a finite sampling approach, which can be described with a dart board analogy. Assume that the point estimate vector is as described in the pie chart of Figure 1A. Assume this pie chart is a dart board. We can throw a finite number of darts at the board (say, 20), which will give us a situation like Figure 1B. The darts are thrown randomly at the board and must all land on the board. We can then calculate the fraction of the 20 darts that land in each wedge of the pie, which gives us a new random vector (shown as new pie chart in Figure 1C). If we repeat this process many times, on average the fraction of darts in each wedge will equal the original point estimate vector. However, any particular throw of 20 darts will likely vary from the original, giving us the random noise that we need.

We can control the amount of variation in the distribution by changing the number of darts that we throw each time. If we throw only 20 darts there is likely to be a fair bit of between the point estimate vector and any particular random vector. However, if we throw 200 darts each time (Figure 1D), each random vector will be relatively close to the original point estimate (Figure 1E). Thus, we can control the amount of variation in our random draws by adjusting what we call the “shape parameter” because it affects the shape of the generator output distribution. If we throw an infinite number of darts, we always get original point estimate vector. We can see the effect of changing this shape parameter by looking at a cumulative frequency plot (Figure 2). As the shape parameter decreases, the range of the random generator distribution increases. This relationship is also illustrated in Figure 3.

This method has the advantage of simultaneously changing all the parameters of the vector, retaining the constraint that they sum to one. A feature of the approach is that the distribution of the random generator for any particular fraction is a function of the value of that fraction. For example, the range of the distribution if the point estimate is 10% for a particular category will be different than the range of a category with a point estimate of 50% (Figure 4). This makes sense if we consider that the distributions are constrained – values can not be less than zero or greater than one so point estimates that are near these boundaries will have different distributions from those of point estimates that are not near the boundaries. Another feature of the method is that fractions that have a point estimate of zero will have value of zero for all random vectors (the width of the pie wedge is infinitely small and no darts can land there). This will not be a problem for most applications.

One limitation of the approach is that the output distribution from the random generator is discrete, rather than continuous. Because we are throwing a finite number of darts (say N), the values in the output random vectors will all be fractions of N (i.e., x/N, where is x an integer and $0 \leq x \leq N$). If, for example, the shape parameter is 20, then there are only 21 possible values in the output vectors, which are 0.00, 0.05, 0.10, 0.15, 0.20, …0.90, 0.95, 1.00 (i.e., values in 5% increments). If the shape parameter was 200, there are 201 possible output values and the output becomes more continuous. The possible values are 0.000, 0.005. 0.010, 0.015, …0.995, 1.000 (i.e., values in 0.5% increments). This discrete output feature is responsible for the "stair step" appearance of the cumulative frequency graphs (Figures 2 and 3). For many applications, the fact that the generator produces only discrete values will not be a problem, but it is useful to be aware of this feature, especially when using shape parameter values less than 20.

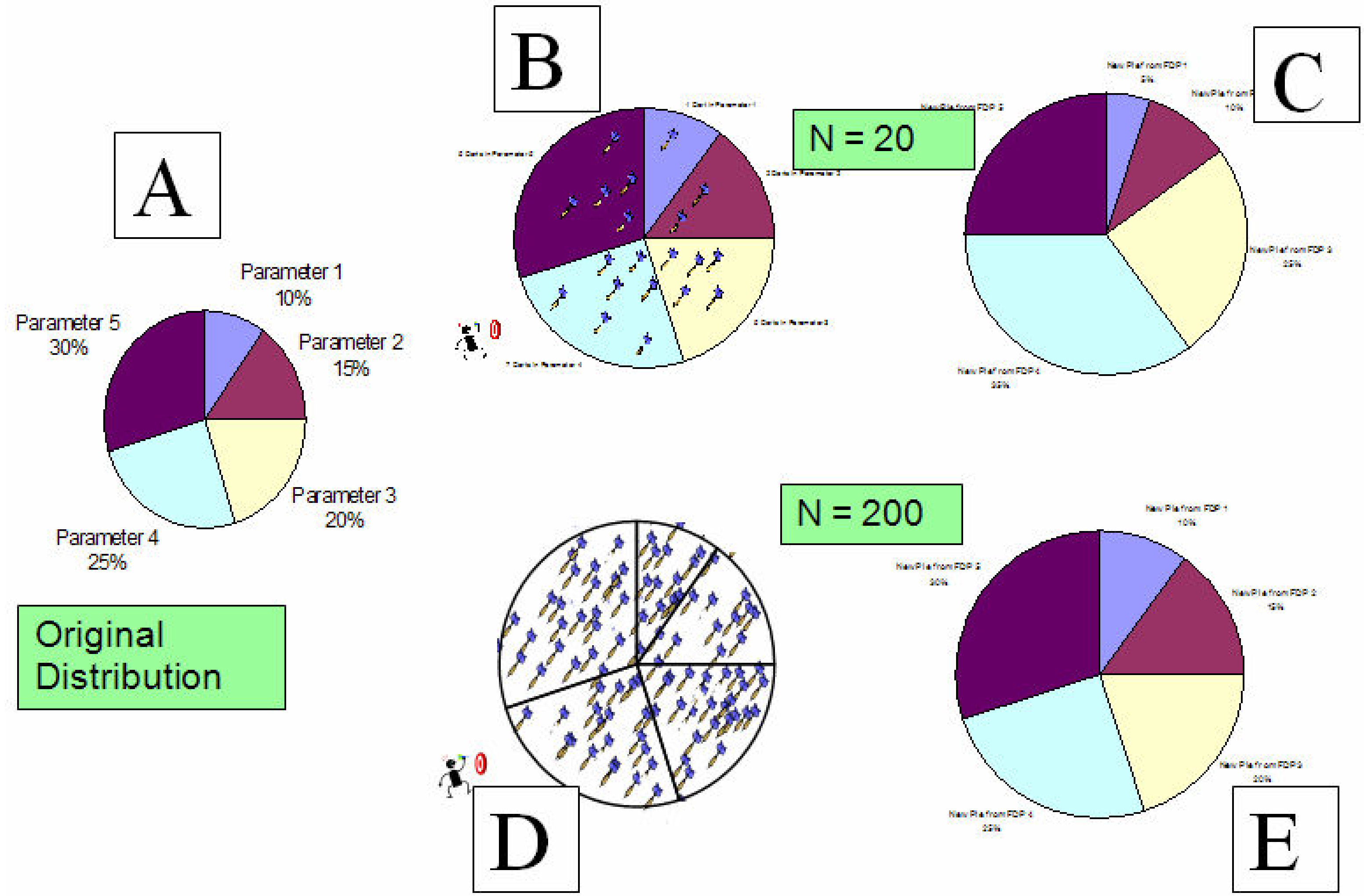

Figure 1: Illustration of multi-nominal finite random sampling method based on a dart board analogy. Pie chart A represents the point estimate fraction vector (a.k.a. dart board). Chart B shows a random sample of 20 darts, with chart C showing the resulting fraction of darts in each wedge, representing the new random fraction vector. Chart D shows a random sample of 200 darts, with chart E showing the resulting fraction of darts in each wedge, representing a new random fraction vector.

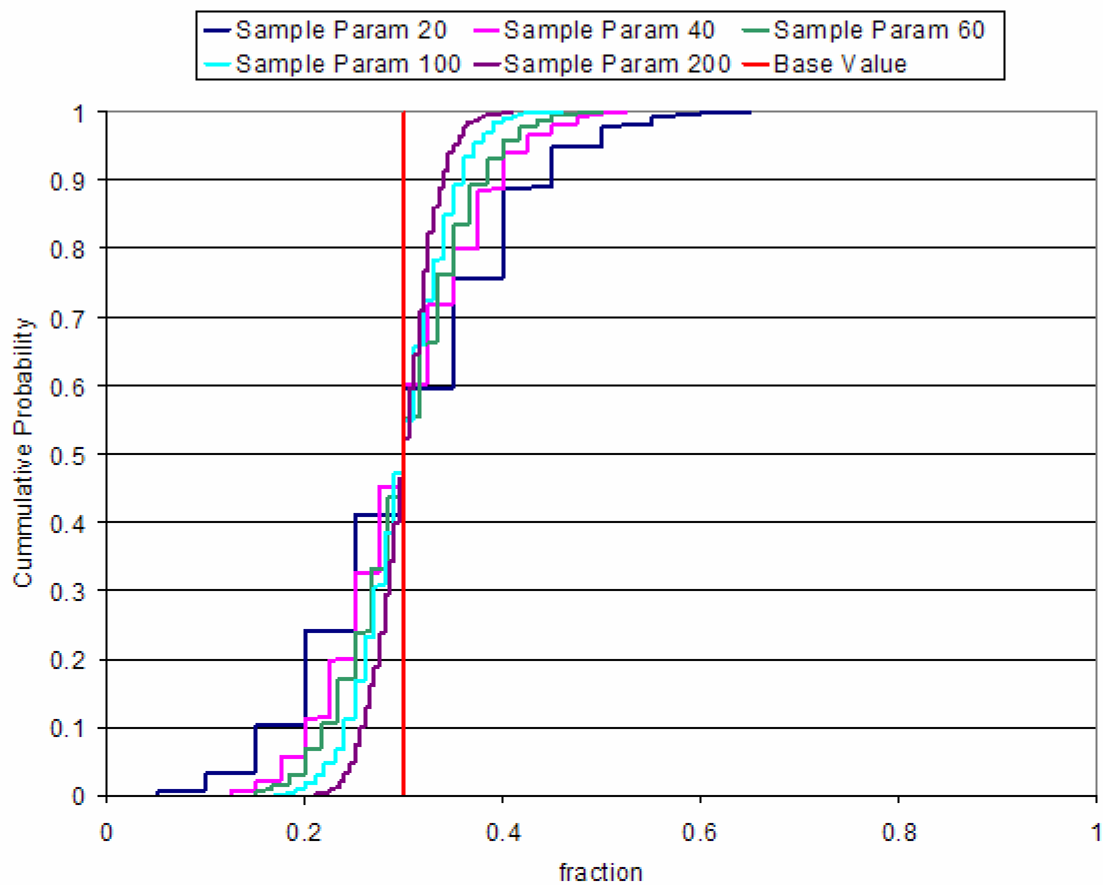

Figure 2: Cumulative frequency distribution of random samples is the point estimate ("Base Value") of 0.3. The different lines indicate distributions with different shape parameters ("Sample Param").

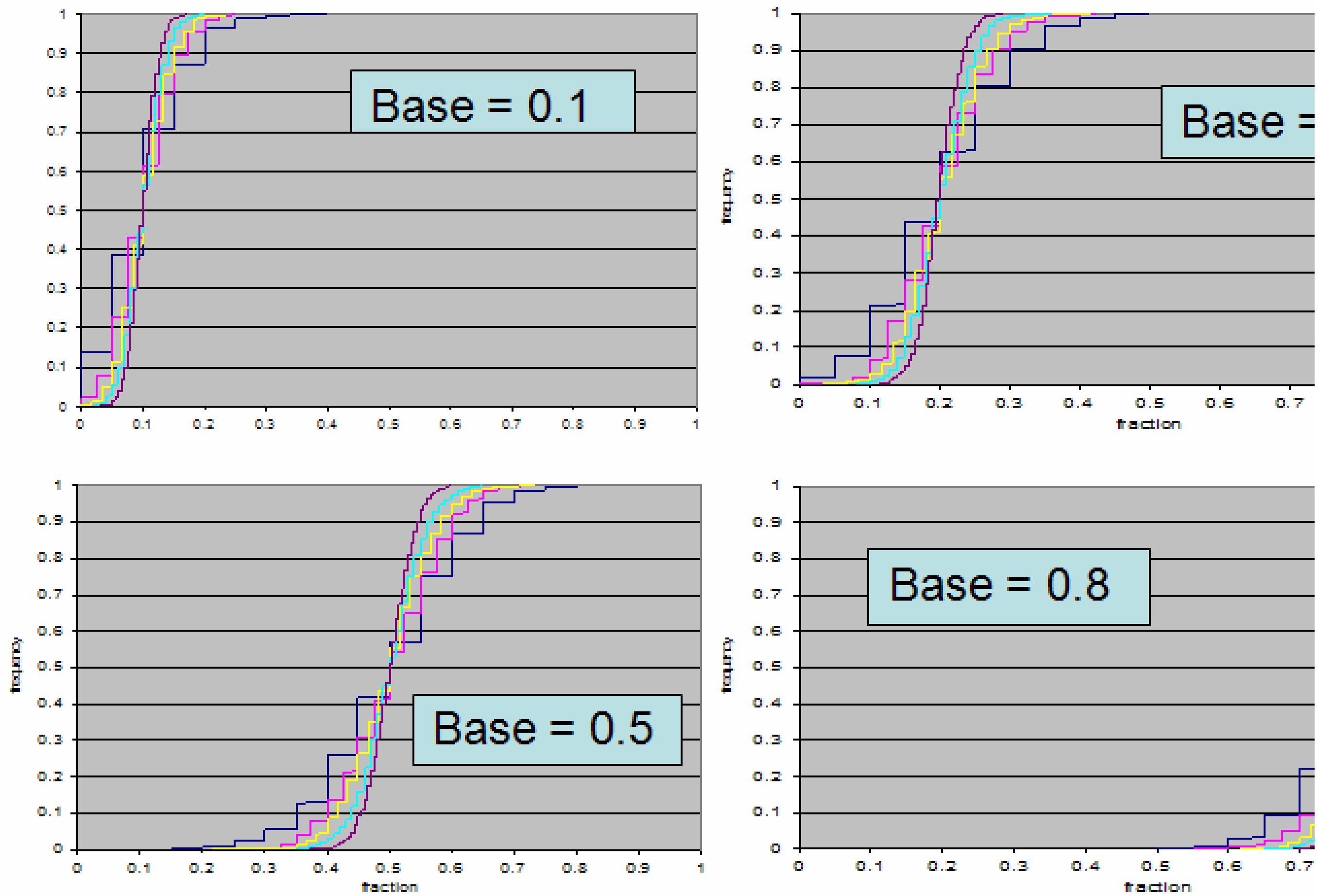

Figure 3: Cumulative frequency distribution of multi-nominal finite random sampling output for different point estimate fractions ("Base"). The different curves are for different shape parameters (values of 20, 40, 60, 100, and 200).

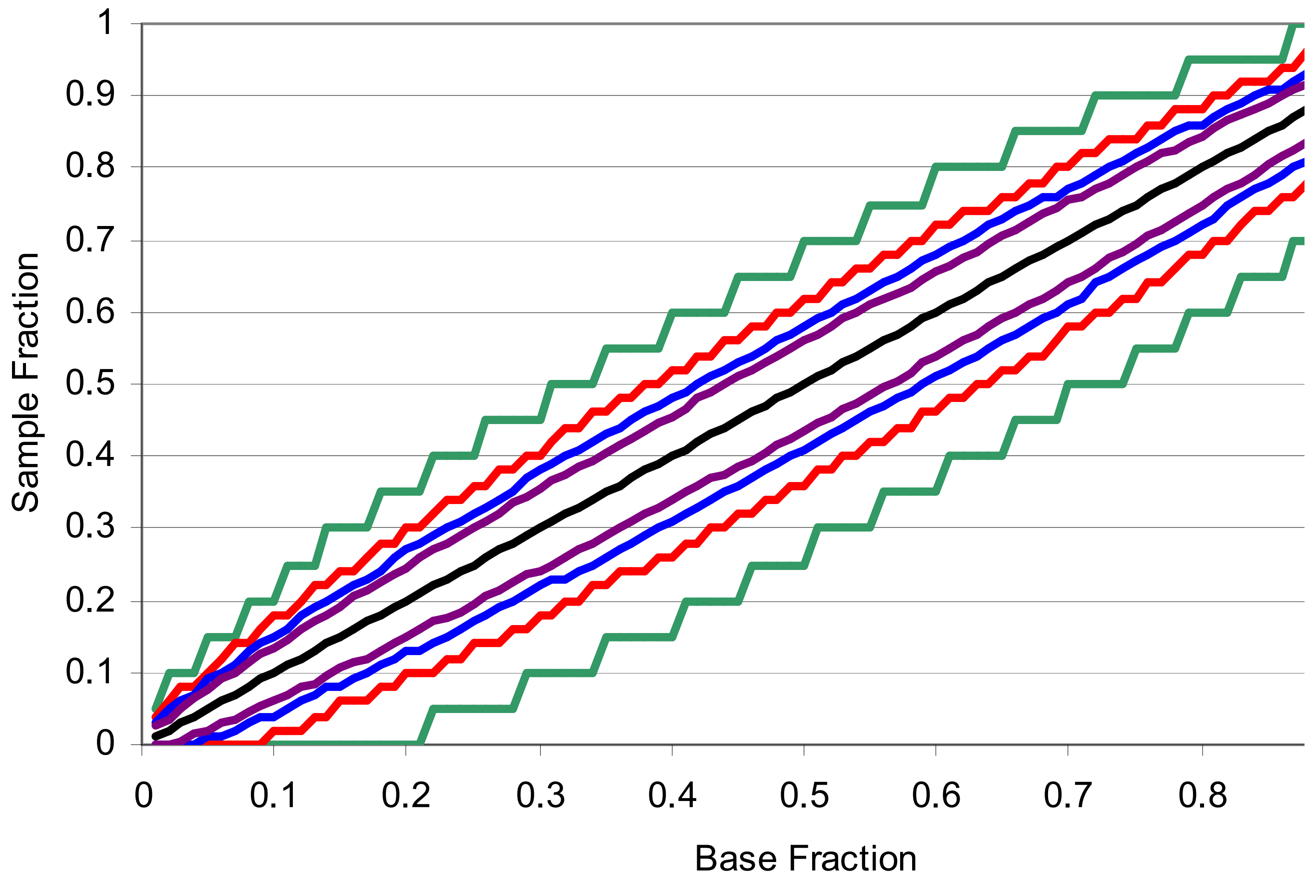

Figure 4: 90% probability bands for multi-nominal finite random sampling output. The black line indicated an infinite shape parameter (the output always equals the point estimate). The purple, blue, red and green lines represent shape parameters of 200, 100, 50, and 20, respectively. To interpret the figure, pick a base fraction (point estimate) on the x-axis then look at the range between the curves on the y-axis at that point. 90% of the output values from the random generator would be within this range.